(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 May 2003 (15.05.2003)

PCT

(10) International Publication Number
WO 03/040994 A2

(51) International Patent Classification[7]: G06K

(21) International Application Number: PCT/US02/34982

(22) International Filing Date:
1 November 2002 (01.11.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/350,117 2 November 2001 (02.11.2001) US

(71) Applicant *(for all designated States except US)*: **ARQULE, INC.** [US/US]; 19 Presidential Way, Woburn, MA 01801 (US).

(72) Inventors; and

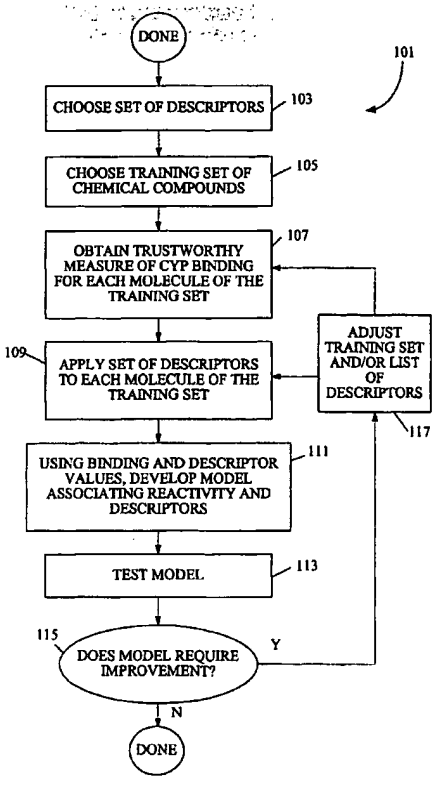
(75) Inventors/Applicants *(for US only)*: **KOCHER, Jean-Pierre** [US/US]; 2645 California Street #212, Mountain View, CA 94040 (US). **EWING, Todd, J., A.** [US/US]; 7730 Yew Court, Newark, CA 94560 (US). **KORZEKWA, Kenneth, R.** [US/US]; 1203 Cristobal Privada, Mountain View, CA 94040 (US).

(74) Agent: **WEAVER, Jeffrey, K.**; Beyer Weaver & Thomas, LLP, P.O. Box 778, Berkeley, CA 94704-0778 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

[Continued on next page]

(54) Title: CYP2C9 BINDING MODELS



(57) Abstract: Computer models described herein predict the binding affinity of compounds with the 2C9 isoform of the Cytochrome p450 family of enzymes. The models predict Ki or pKi for arbitrary compounds. They accomplish this by using selected molecular properties of the compound in question. Numeric values of these properties (also called descriptors) are received by the model. Execution of the model performs a calculation based on these descriptors and returns a value of binding affinity. The descriptors may include a measure of lipophilicity, a measure of aromaticity, and a measure of partial negative charge on the compound.

WO 03/040994 A2

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

CYP2C9 BINDING MODELS

Jean-Pierre Kocher, Todd J.A. Ewing and Kenneth R. Korzekwa

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority from US Provisional Patent application no. 60/350,117, filed November 2, 2001, naming Kocher et al. as inventors, and titled "2C9 Binding Affinity Modeling." That provisional application is incorporated herein by reference for all purposes.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

The U.S. Government may have certain rights in this invention pursuant to NIH Grant No. 1R43 GM63427-01.

FIELD OF THE INVENTION

The present invention relates generally to methods, apparatus, and program products for predicting binding of compounds to cytochrome p450 enzymes from descriptors of physicochemical properties. The invention also relates to methods, apparatus, and program products for generating models that predict activity from such descriptors.

BACKGROUND

Drug development is an extremely expensive and lengthy process. The cost of bringing a single drug through the research and development pipeline to market is about $500 million to $1 billion dollars, with average time duration from about 8 to 15 years. Drug development typically involves the identification of 1000 to 100,000 candidate compounds that eventually lead to a few marketable drugs.

New chemical entities (NCEs) enter the drug discovery pipeline through a rational drug design process in which information about the target of action is used to

design the lead compound. The NCE design process alone necessitates an average time investment of six years, with an estimated cost of more than $30 million per molecule. More than 90% of drug candidates that enter clinical development fail prior to making it to commercial launch.

In the traditional drug discovery process, physicochemical properties and ADMET/PK (absorption, distribution, metabolism, elimination, and toxicity/pharmacokinetics) parameters are evaluated, extending the time required to identify a lead candidate. Even optimized leads that have passed these evaluations may ultimately show undesirable ADMET/PK properties, necessitating abandonment or re-design which contributes greatly to the cost of drug development. Overall, ADMET deficiencies account for 50-60% of compound failures during early development. ADMET considerations are also responsible for the vast majority of drug-drug interactions. Thus it would be highly desirable to optimize the selection of candidate compounds by being able to predict the correlation between ADMET properties and the physicochemical structure early in the course of molecular design. Such information can be used early in development to filter out certain candidates that might otherwise get much further and consume significant resources. In addition, the information can be used to flag other candidates for fast track development given a prediction of desired ADMET characteristics. Still further, such information may guide redesign or "rescue" efforts for compounds that would show great promise but for one or more ADMET problems. Redesign involves some chemical structural modification (usually relatively minor in comparison overall complexity of the molecule) intended to improve ADMET performance. For example a very labile moiety on a compound deemed to metabolize too fast could be replaced with a more stable moiety.

A large portion of all drug metabolism in humans is carried out in the liver by the cytochrome P450 enzyme system. The cytochrome P450 enzymes (CYP) are a superfamily of heme-containing oxidase enzymes that are involved in the metabolism of hydrophobic drugs, carcinogens, toxic compounds, and metabolites circulating in the blood. There are numerous subfamilies, often termed "isozymes" or "isoforms." The most important CYP enzymes in drug metabolism are the CYP3A4, CYP2D6 and CYP2C9 isozymes. It is estimated that in humans, 50% of all drugs currently on the market are metabolized by the cytochrome P450 family, with CYP2C9 alone responsible for the metabolism of approximately 20% of marketable drugs.

A knowledge of the binding characteristics of an exogenous compound to a metabolic enzyme is necessary in the understanding of the metabolism of that

compound. Numerous features of the substrate can modulate binding affinity. These include the presence or absence of positive charges, negative charges, hydrogen bond donors, hydrogen bond acceptors, aromatic centers, and hydrophobic centers, and various whole molecular characteristics such as size, partition coefficient, etc.

CYP2C9 is responsible for a significant percentage of the drug metabolism of the cytochrome P450 system, which is due to the fact that the CYP2C9 subfamily has the capacity to metabolize multiple substrates. Some substrates can bind tightly to the enzyme leading to a reduction in the metabolism of other substrates; hence, a propensity for competitive inhibition exists, which often results in serious toxicity and concomitant adverse drug interactions. As no structures of the P450 enzymes have been solved, structural-based computational approaches cannot be applied to elucidate binding affinity of CYP2C9 to substrates. Therefore, new methods are needed to build new models for predicting CYP2C9 binding affinity. Such methods and models are described herein.

SUMMARY

This invention addresses these needs by providing methods and computer models for predicting binding affinity of compounds with the 2C9 isoform of the Cytochrome p450 family of enzymes (CYP2C9). The models predict absolute or relative values of binding affinity, e.g., Ki or pKi for arbitrary compounds. They accomplish this by using selected molecular properties of the compound in question. Numeric values of these properties (also called descriptors) are received by the model. Execution of the model performs a calculation based on these descriptors and returns a value of binding affinity. Preferably, the descriptors include at least a measure of lipophilicity, a measure of aromaticity, a measure of size, and a measure of partial negative charge on the compound.

One aspect of this invention pertains to methods of predicting binding of a compound to CYP2C9. The method may be characterized by the following sequence: (a) receiving a value representing the lipophilicity of the compound; (b) receiving a value representing a negative charge or partial negative charge associated the compound; (c) receiving a value representing the aromatic character of the compound; and (d) calculating the binding of the compound to CYP2C9 with an expression treating the values received in (a)-(c) as independent variables and treating the binding to CYP2C9 as a dependent variable. Other independent variables for the expression may include the size of the compound, and the flexibility of the compound

(as indicated by number of rotatable bonds for example). In one embodiment, the calculated binding is provided as a value of Ki or pKi.

The expression may take many different forms. Often, it will have a very simple form that requires very little computational effort to execute. In one example, the expression takes the form of a linear relationship between said independent variables (descriptors) and the binding to CYP2C9. In a specific preferred embodiment, the expression includes at least the following independent variables: number of aromatic atoms, molecular weight, number of rotatable bonds, a partitioning property, and a surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

The descriptor values received in (a)-(c) may take many forms. For example, the value representing the lipophilicity may be a measured or predicted value of a partitioning property, such as a partition coefficient or distribution coefficient. Alternatively, or in addition, the value representing the lipophilicity may be a measure or prediction of the hydrophobicity of the compound, as represented by the number of hydrophobic atoms in the compound or the surface area of the compound occupied by hydrophobic atoms, for example. Further, the value representing the partial negative charge on the compound may be, for example, the surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level (e.g., < -0.2). Still further, the value representing the aromatic character may be one or more of (i) the number of aromatic atoms in the compound, (ii) the type of aromatic atoms in the compound, (iii) the number of aromatic rings in the compound, and (iv) the type of aromatic rings in the compound.

The methods described above may be performed on a number of different compounds. Some of these compounds are selected because they have a calculated binding to CYP2C9 that exceeds a predefined value. Depending on the magnitude of that predefined value, the selected compounds may be characterized as potential inhibitors having drug interference difficulties. Alternatively, they may be evaluated by a computer model that predicts reactivity of specific chemical sites. In the later case, the binding expression serves as a filter that prevents non-binding compounds from being analyzed further. This methodology is premised on the fact that compounds that cannot bind to CYP2C9 cannot be metabolized by CYP2C9.

Another aspect of the invention pertains to computer-implemented methods of creating a multivariate model for predicting the binding of compounds to CYP2C9. The method may be characterized by the following operations: (a) for each compound in a training set, receiving values representing (i) the binding of the compound to

CYP2C9 and (ii) certain molecular descriptors; and (b) fitting the values to create the multivariate model of binding to CYP2C9. In these methods, the molecular descriptors include at least the lipophilicity of the compound, a measure of the negative charge or partial negative charge on the compound, and the aromatic character of the compound. Thus, the model of binding to CYP2C9 is also as a function of these descriptors. Addition of other descriptors to the model can be provided as described above. In a preferred embodiment, the fitting involves performing a regression a data set comprising values of lipophilicity, partial negative charge, aromatic character, and binding to CYP2C9 for various compounds comprising a training set.

Yet another aspect of the invention pertains to computer program products including machine-readable media on which are provided program instructions for implementing the methods described above, in whole or in part. Frequently, the program instructions are provided as code for performing certain method operations. Any of the methods of this invention may be represented, in whole or in part, as program instructions that can be provided on such machine-readable media. In addition, the invention pertains to various combinations and arrangements of data generated and/or used as described herein.

These and other features of the present invention will be described in more detail below in the detailed description of the invention and in conjunction with the following figures.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a high-level flowchart depicting typical operations that may be employed to generate a model in accordance with an embodiment of this invention.

Figure 2 is high-level flowchart depicting methods for predicting the binding and metabolic rate of a substrate molecule, starting with the specified descriptors of the substrate.

Figure 3 is a block diagram of an Internet based system for analyzing therapeutic compounds in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT

A. INTRODUCTION

The present invention pertains to methods, apparatus, and program code that use simple, rapidly executing models to predict the binding of arbitrary compounds to CYP2C9. In one approach, the model can predict Ki or pKi values (pKi = -log Ki) for binding of compounds to CYP2C9. Alternatively, the model can generally "classify" compounds as binding or non-binding.

In the models, CYP2C9 binding affinity is typically the dependent variable and various molecular descriptors are the independent variables. In a preferred embodiment, the molecular descriptors include a combination of some or all of the following: lipophilicity (e.g., logP), size (e.g., molecular weight), surface properties (e.g., accessible surface area of certain types of atoms such as polar, negative, or hydrophobic atoms), nature of atoms (e.g., aromatic), negative charges or partial negative charges on the compound, and flexibility (e.g., number of rotatable bonds).

In a more specific embodiment, the descriptors relevant to CYP2C9 binding affinity include logP (partition coefficient), number of aromatic atoms, number of hydrophobic atoms, van der Waals accessible surface area of hydrophobic atoms, molecular weight, number of heavy atoms, total number of atoms, number of rotatable bonds (non-aromatic bond, non-double bond, etc.), number of hydrogen bond donors, and accessible surface area of atoms having less than a defined negative partial charge (e.g., less than –0.2).

The models of this invention have various uses. For example, they may predict compounds that are likely to bind very tightly to the CYP2C9 enzyme. Such compounds may be flagged as potential inhibitors that could result in a drug interference problem. Alternatively, the binding models of this invention could be used in conjunction with other models to predict rates of metabolism by CYP2C9. Only compounds that can bind to CYP2C9 would be expected to be metabolized by this enzyme. Hence, the model can serve as a filter to exclude non-binding compounds from further analysis by other software that predicts actual rates of metabolism for binding compounds and other software that predict regioselectivity of CYP2C9 for binding compounds. Since compounds can be metabolized by enzymes other than CYP2C9, this model may be used in conjunction with other models that predict binding to non-CYP2C9 enzymes (e.g., CYP3A4, CYP2D6, etc.).

The CYP2C subfamily of P450 enzymes is of major importance in drug metabolism. The most abundant 2C isozyme, CYP2C9, accounts for approximately 17-20% of the total P450 content in the human liver. CYP2C9 is a monooxygenase enzyme, and exhibits high regioselectivity and a broad substrate specificity for negatively charged substrates. This specificity has been rationalized in terms of a hydrogen bond donor/acceptor model. CYP2C9 acts via a hydroxylation of its substrates.

CYP2C9 is known to be membrane-bound protein with an approximate molecular weight of 50 kD, and is mainly found in the endoplasmic reticulum of hepatocytes. However, the structure of CYP2C9 has not been definitively elucidated, as membrane-bound proteins do not readily yield crystals. CYP2C9 is believed to catalyze mostly oxidative reactions such as hydroxylations, requiring a specific electron donor NADPH and O_2 in stoichiometric amounts to the hydroxy product formed.

A limited amount of information exists regarding the binding characteristics of the CYP2C9 isozyme. It is generally known that the binding of substrates to the active site largely relies on hydrophobic interactions. Further research has hypothesized that a hydrogen bond donor in the active site of CYP2C9 exists which dictates substrate orientation (Jones, et al., www.joneslab.wsu.edu). At this point, to assist in understanding the concepts presented herein, explanations of some pertinent terms are provided. The scope of the invention should not necessarily be limited by the following examples.

"Physicochemical property" (or sometimes just "property") refers to a particular physical and/or chemical property of a compound under consideration. Multiple physicochemical properties are employed by models of this invention to predict binding to CYP2C9. In preferred embodiments of this invention, the properties pertain to the compound as a whole. Hence they are sometimes referred to as "molecular" properties in order to distinguish them from properties pertaining only to a specific region or fragment of the compound, or even to individual atoms within the compound. Examples of whole compound physicochemical properties include partition coefficient (P), molecular weight (MW), formal charge (FC), total van der Waals surface area associated with various types of atoms, total number of hydrogen bond donors/acceptors, etc.

Note that in alternative embodiments, a model of this invention may make use of one or more region specific or atom specific properties. Examples of region specific physicochemical properties include pre-calculated properties (e.g.,

hydrophobicity) of generic molecular fragments obtained from molecules using fragmentation rules. Examples of atom specific physicochemical properties include chemical information about a site atom such as information about its neighbor atoms, its partial charge, its total charge, bond lengths, whether it is a hydrogen bond donor or acceptor, etc. Further examples will be set forth below.

For purposes of this invention, particularly relevant physicochemical properties are those found to impact the binding to CYP2C9. For example, the hydrophocity, negatively charged regions and size of a compound often have a pronounced effect on the binding affinity of a compound to CYP2C9.

The term "descriptor" refers to a variable or value representing a property of a particular compound. Thus, the term is closely related to, and in a sense overlaps with, "physicochemical property." Descriptors may be viewed as quantitative or textual representations of properties. They appear as independent variables in expressions or models for predicting "binding" of a particular compound. A potentially infinite number of descriptors may characterize a compound. Multivariate models employ two or more descriptors to predict the binding of a compound.

The descriptors may be two-dimensional or three-dimensional. Two-dimensional descriptors can be calculated from the atoms, connections or bonds between atoms. They are not dependent on atomic coordinates or molecular conformations. Examples include elements identification, atom counts, formal charges, bond orders, etc. Three-dimensional descriptors depend on internal coordinates and/or absolute orientation. One examples is a dipole moment. In certain preferred embodiments of this invention, only two-dimensional descriptors are employed.

"Binding" refers to the ability of a compound (commonly referred to as a ligand) to combine with an enzyme or other protein (a CYP enzyme in the case of this invention). Enzymes typically have "binding site" where the compound has the greatest affinity. Here the compound-enzyme complex exists in a low energy resulting from the stabilizing influence of hydrogen bonding or other physicochemical interaction. In a sense, binding is a "property" of a compound. However, in the context of this invention, the binding of a compound is usually a characteristic that is to be predicted. In other words, binding serves as a dependent variable related to descriptors, which are independent variables. Depending on how a model is constructed, binding may take the form of a specific numerical value (e.g., an equilibrium constant such as K_i) or a threshold or filter (e.g., binds or does not bind).

A "Model" is a mathematical or logical representation of a physical and/or chemical relationship. Models of this invention predict compound-CYP binding from one or more descriptors of physical and/or chemical properties. As just mentioned, such models treat binding as a dependent variable and descriptors as independent variables.

Models can take many different forms. They can take a very simple format such as a look up table or a more complex format such as a molecular simulation or docking algorithm. Examples of the mathematical/logical form of models include linear and non-linear mathematical expressions, look up tables, neural networks, support vector machines, Bayesian models, classification and regression trees/graphs, clustering approaches, and the like. In one preferred embodiment, the model form is a linear additive model in which the products of coefficients and descriptors are summed. In another preferred embodiment, the model form is a non-linear product of various descriptors.

Models can predict binding as a discrete event or a continuous range. A classification model predicts whether or not a discrete event such as binding will occur. A continuously variable model will predict the probability that the event will occur or the strength of the event (e.g., K_i for enzyme substrate binding).

Models are typically developed from a training set of chemical compounds or other entities that provide a good representation of the underlying physical/chemical relationship to be modeled. Together, the activities and descriptors of compounds form members of the training set and are used to develop the mathematical/logical relationship between binding activity and descriptors. This relationship is typically validated prior to use for predicting activity of new compounds.

"Fitting" refers to the act of mapping a set of data points (a data set) to a mathematically or logically convenient format. That format is frequently a mathematical expression of linear or non-linear form. Look up tables may also be used if the values in the tables represent consensus values obtained from the data set. The expression (or other logical representation) resulting from the fit data generalizes the data in manner that can be used to predict activity from compounds or other entities outside the data set. Thus, the expression may be viewed as a "model." Generally, fitting techniques are referred to as optimization or minimization techniques. Specific examples of fitting techniques include Newton's method, various splines (e.g., cubic splines), least squares, partial least squares, various other regression techniques, as well as simplex, Monte Carlo, Tabu, and genetic algorithm optimization techniques.

B. GENERATING A MODEL FOR APPROXIMATING CYP2C9 BINDING

1. OVERVIEW

Figure 1 presents a process flow diagram depicting typical operations that may be employed to generate a model in accordance with an embodiment of this invention. As depicted, a process 101 begins with the choice of an appropriate set of molecular descriptors for characterizing organic molecules that may bind with CYP2C9. See 103. The descriptors should correlate with CYP2C9 binding.

With the associated descriptors chosen, the next process operation identifies an appropriate training set of organic molecules. See 105. These molecules are chosen to provide a significant sampling of the types of structural characteristics and CYP2C9 binding affinities that the model is likely to encounter in practice. For each member of the training set, the process receives trustworthy values of binding to CYP2C9. See 107. Typically, these values are obtained experimentally, although in the case of a very reliable and detailed model, they may be obtained computationally. This later situation might be appropriate where the model is too slow or otherwise requires too many computational resources to be useful for quickly analyzing large numbers of compounds that a research organization might present. A much faster or more efficient model would be required for this analysis. The faster model could be generated from the training set comprised of some data values obtained from the more reliable, albeit inefficient, model.

The experimental binding affinity constitutes one component of each data point used to construct the models of this invention. The other component is the descriptor values. Using the set of descriptors identified at 103, the process receives actual values of those descriptors for each member of the training set. See 109. For example, one descriptor may be the van der Waals surface area associated with hydrophobic atoms on the compound. The procedure may obtain these descriptor values by analyzing the simple two-dimensional or three-dimensional chemical structures of the members of the training set.

Once the descriptor values have been calculated, each member of the training set is now represented by a set of descriptor values and a trustworthy measure of CYP binding. Then, using these data points, the process generates the actual model that associates reactivity with the descriptors. See 111. The model may take the form of a

simple expression including coefficients for each descriptor value. A more detailed example of a model generating process will be described below.

With the model in hand, the methodology 101 may test the model against a particular test set of molecules (or some actual field test molecules). See 113. The molecules used in the test should have known binding values for the CYP isoform under consideration (e.g., CYP2C9). The degree to which the model accurately predicts binding determines whether it needs improvement. See 115. Assuming that the model does a good job of predicting CYP binding, process 101 is complete. Assuming that the model needs improvement, then a revised training set or list of descriptors is chosen. See 117. From there, process control returns to 107 or 109 as appropriate. The revised set or list is chosen to handle the types of molecules or structural features that presented difficulty to the model.

2. CHOOSING A SET OF DESCRIPTORS

As indicated above, the models of this invention make use of specific descriptors for chemical compounds, particularly organic molecules. These descriptors represent properties that should affect binding to CYP2C9. Particularly interesting descriptors will be described in more detail below.

Any organic molecule under consideration, whether used in a training set or an investigation set, is characterized using an appropriate set of descriptors. The descriptor characterization of the molecule is then used to either generate a model (the molecule is part of a training set) or predict binding (the molecule is part of an investigation or test set).

. At some point in the model development process, one must select a set of descriptors to represent the activity in question. Experience has shown that some of the most useful descriptors for CYP2C9 binding affinity include partitioning properties such as logP and logD and SlogP, hydrogen-bond donor/acceptor counts; size based descriptors (e.g., molecular weight, calculated molar refractivity, atom counts, computed surface area, and computed volume), surface area of hydrophobic atoms and atoms exhibiting a partial negative charge, flexibility, and the like. These and other descriptors are described in the documentation provided with the software MOE 2001.01, QuaSAR-Descriptor available from Chemical Computing Group Inc. of Montreal, Quebec. This documentation is incorporated herein by reference for all purposes.

For CYP2C9, the following descriptors have been found to correlate with binding: aromatic character, lipophilicity (or hydrophobicity), negative charge or partial negative charge, flexibility, size, and hydrogen bond donating/accepting capabilities. Such descriptors may be used in various forms. The following discussion provides a brief explanation of each such descriptor.

The "aromatic character" of a compound generally refers to some quantifiable measure of the aromaticity in a compound. As generally understood in the art, aromatic groups molecular components associated with planar cyclic conjugated systems of double and single bonds having delocalized π electrons. Generally, they undergo electrophilic substitution. Examples include compounds that contain a benzene ring, certain heterocyclic compounds such as pyridine compounds, imidazole compounds, thiophene, etc., as well as non-benzenoid, non-heterocyclic compounds such as ferrocene, azulene, and tropylium cation containing compounds.

The aromatic character of such compounds may be quantified by the number and type of aromatic atoms, the number and type of aromatic rings (e.g., fused ring systems, heterocyclic systems, etc.), number of aromatic bonds, etc. The "type" of aromatic atom may simply be the element identification (N, C, S, etc.) or it may be something more complex that accounts for its neighborhood within the substrate (e.g., no atoms within one bond are nitrogens). The "type" of aromatic ring may specify ring sizes, fused ring systems, heterocyclic rings, benzene rings, etc. Any of these may represent a descriptor of the compound in question. For each compound, the molecular descriptor is given a particular numeric value (e.g., 5 aromatic atoms or 1 fused ring system) or a particular logical value (e.g., has or does not have a benzene ring).

"Lipophilicity" is another molecular descriptor. It is generally used to describe the transport processes of a compound in a biological system, and represents the ability of a drug or entity to dissolve in a lipid phase when an aqueous phase is also present. Lipophilicity is a major structural factor that influences the pharmacokinetic and pharmacodynamic behavior of compounds. Partitioning within a biological system and biological activity are governed by forces that are defined by hydrophobic interactions.

The term "hydrophobicity" is often used interchangeably with lipophilicity, but refers more specifically to interaction properties on molecular surfaces. Hydrophobicity is the association of nonpolar groups in an aqueous environment, which arises from the tendency of water to exclude nonpolar molecules. Strong hydrophobic interactions can result in non-specific binding with proteins in the

aqueous environment. A hydrophobic drug molecule has a tendency to reduce the surface area exposed to water; hydrophobic compounds will therefore tend to bind to hydrophobic surfaces through Van der Waals bonds.

Hydrophobicity may be characterized in many different ways. Some of the simple characterizations include the number of hydrophobic atoms in a substrate and the substrate surface area occupied by hydrophobic atoms. The most widely used measure for the lipophilic/hydrophobic properties of a compound in many biological partitioning processes is the logarithm of the equilibrium concentration in octanol-water, logP, which has been established as a quantitative parameter for the lipophilicity of a compound. The properties of 1-octanol are thought to resemble those of lipid bilayer membranes, and it is believed that distribution of chemicals into 1-octanol simulates their ability to passively diffuse across biological membranes.

More generally, one can refer to a "partitioning property" that characterizes the ability of a compound to partition between two immiscible phases: one aqueous and the other non-aqueous. Various types of partition coefficients are defined. The variations arise for the most part from the choice of non-aqueous phase and in the amount of buffering employed. The non-aqueous phase is, as indicated, typically 1-octanol. The aqueous phase is a phosphate buffer of 7.4 pH.

A widely used variation of the partition coefficient is the pH dependent partition coefficient (distribution coefficient, D), typically presented as logD. This is typically measured experimentally in a similar manner to logP except that the pH of the aqueous buffer is altered, and measurements taken over a pH range (1-14). Alternatively, a calculated logD may be obtained, but this requires additional knowledge or calculation of the dissociation constants associated with all the ionizable groups within a given molecule. Due to this, calculation of logD is often less reliable than that of logP.

SlogP, which is a model that attempts to predict the experimentally derived logP value, is another widely used partition coefficient. It is the logarithm of a calculated octanol/water partition coefficient, which includes implicit hydrogen atoms. The calculation is described in Wildman and Crippen, Prediction of Physiochemical Parameters by Atomic Contributions, J. Chem Inf. Comput. Sci. 39: No. 5, 868-873 (1999). SlogP has been used as a surface area descriptor that is intended to reflect the hydrophobic and hydrophilic effects in a receptorPartition coefficients can be easily measured, but must be done in a consistent manner. As indicated, certain accepted models can predict/calculate the partition coefficients of

arbitrary compounds. Typically, though not necessarily, descriptors based on a partitioning property are provided as the logarithm of a partition coefficient.

The "negative charge" or "partial negative charge" associated with a compound refers to a characteristic of whole compound, or a specific part of the compound, which has the tendency to attract positive charge. Note that the binding pocket of CYP2C9 is believed have a positively charged region, and thereby preferentially bind negatively charged regions of a substrate. An ionized compound will have a formal charge that is some whole number (e.g., -1, -3, etc.). A region of a compound may have a partial negative charge where it deviates from charge neutrality and has a high electron density. In a molecule, the formation of bonds lead to a redistribution of the valence electron density, and this may result in regions where there is an imbalance between the ion core charge and the immediately-surrounding valence electron charge. Often a negative partial charge results at atoms having strongly electron withdrawing properties.

Examples of descriptors representing the negative charge or partial negative charge on compounds include the formal charge of the compound, the number of atoms having a partial negative charge of a magnitude greater than a defined level, the total surface area associated with such atoms, and the like. In a particularly preferred embodiment, the relevant descriptor is the sum of the van der Waals surface areas of all atoms having a partial negative charge of –0.2 or lower.

"Molecular flexibility" is a topographical parameter, which is an indicator of the number of rotatable bonds in a molecule. Examples of molecular flexibility parameters include the total count of rotatable bonds in the molecule and the fraction of bonds in the molecule that are rotatable. A rotatable bond is defined as any single non-ring bond, bound to a nonterminal nonhydrogen atom.

Another useful class of descriptors is the "size-based" descriptors. These include, but are not limited to, molecular weight, atom count, heavy atom count (e.g., number of atoms having an atomic weight of greater than a predetermined weight such as 1), molecular surface area, and McGowans volume (Vx) to name just a few.

Another useful class of descriptors for CYP2C9 binding affinity is the various molecular surface area descriptors, which specify the molecular surface area that exhibits particular properties . Examples include total molecular surface area exhibiting partial negative charge of magnitude greater than about –0.2, total molecular surface area occupied by polar groups, total molecular surface area occupied by acidic or basic groups, etc.

A "hydrogen bond donor" is generally defined as a hydrogen atom covalently bound to an electronegative atom or group such as an oxygen atom, a sulfur atom, or a nitrogen atom. Examples of hydrogen bond donors include the hydrogen atoms in hydroxyl groups, carboxylate groups, amides, amines, mercaptans, and the like. A "hydrogen bond acceptor" on a compound is generally defined as an electronegative atom having a free electron pair. Often oxygen, nitrogen, and sulfur heteroatoms serve as hydrogen bond acceptors. Examples include oxygen atoms, disulfide sulfur atoms, and amine nitrogen atoms (in primary, secondary, or tertiary amines). Lipinski's rule of five suggests that most drug-like compounds have less than ten hydrogen bond donors and/or hydrogen bond acceptors.

A skilled chemist can quickly identify hydrogen bond donors and acceptors on an arbitrary chemical compound. Similarly, numerous computation tools can make this assessment. Examples include Cerius 2 (Accelerys Inc), MOE (Chemical Computing Group Inc).

3. CHOOSING A TRAINING SET

Materials for which models of this invention may predict a pertinent activity include most any compound introduced (such as by ingestion or inhalation) into a living organism. Particularly preferred compounds for analysis are potential therapeutics considered in a drug discovery effort. In developing a model of CYP binding for such compounds, one should carefully choose a training set. A large group of structurally diverse chemical compounds should be used. Generally, a training set member may be any compound that has been synthesized and has had its binding to an appropriate CYP enzyme measured. Alternatively, some members of a training set could be rendered virtually, but only if their binding affinity for CYP has been characterized by a proven reliable computational technique.

The specific compounds chosen for the training set may also be focused on the chemical structural space relevant to the model. Thus, if a model is to be developed for potential therapeutic compounds taken orally, then the training set should include various small organic drug-like molecules. For example, it may be unnecessary to consider compounds having a molecular weight greater than about 1000, because these compounds are unlikely to be metabolized by a CYP enzyme.

Often distinct training sets are used for developing separate types of models. Model examples include binding to CYP3A4, binding to CYP2D6, and binding to

CYP2C9. The training set for binding to CYP3A4 should be diverse in the types of moieties relevant to CYP3A4 binding. Similarly, training sets for binding to CYP2D9 should be diverse in the types of moieties relevant to CYP2D9 binding.

The diversity in the training set should reside in the descriptors of interest. Such diversity may be manifest in a wide range of "scaffolds" and "building blocks" (e.g., a wide range of ring systems, substitutions, etc.)

In an approach employed to generate the model depicted in the specific equation for pKi (for CYP2C9 binding) shown below, members of a training set were selected by the following set of descriptors: formal charge, total positive van der Waals surface area, total negative van der Waals surface area, logP, aromatic atom count, and non-hydrogen atom countA principal component analysis was performed on the points in space defined by the descriptor combinations of each compound. A "chemical space" was defined by the highest order principal components. Each potential member of the training set was placed at its position in this space. Then a subset of the potential members was selected to represent the full expanse of the chemical space without providing too much redundancy. The selected compounds were then analyzed for pKi for CYP2C9 binding and found to have a relatively even distribution from about 3 to 7.

4. GENERATING A MODEL OF CYP2C9 BINDING

When the appropriate training set has been selected and characterized by binding to CYP2C9 and pertinent descriptors, then the model can be generated by an appropriate data fitting technique. Associating binding with particular descriptors generates the model. Generally, "association" represents an attempt to find a relationship between the two groups of variables. Examples of data fitting techniques that may be used in embodiments of this invention include various regression techniques, partial least squares, back-propagation neural networks, linear discriminant analysis and genetic algorithms.

A linear regression equation relates independent and dependent variables ($Y = XB+e$ where Y is the dependent variable represented by a vector (i.e., reactivity of site of the training set members), X is the independent variable represented by a matrix (i.e., structural descriptors grouped by training set members), B is the regression coefficient represented by a vector, and e is the residual). PLS (Projection to Latent Structures or Partial Least Squares) regression analysis is most commonly

used with this invention because it can process large numbers of correlating descriptors while minimizing the risk of over-fitting.

C. USING THE MODEL TO PREDICT BINDING

One aspect of this invention pertains to using methods and models for predicting binding affinity of compounds. Such methods may be characterized as follows. First, the implementing system identifies the chemical compound in question. Second, it identifies values for one or more descriptors of the compound. These are the descriptors used in the method/model. Third, the system combines the descriptor values for the compound in question (in the manner required by the model format) to predict the binding affinity of the compound in question. Finally, the system outputs a calculated binding value for the chemical compound. The system may display the calculated activity value for the compound, on a computer display screen or output medium, for example.

The models of this invention predict binding to a particular metabolizing enzyme such as a CYP enzyme. Such binding model may represent a consensus binding to multiple genetic isoforms of an enzyme, or, more preferably, binding to a specific isoform of an enzyme (CYP2C9, in a particularly preferred embodiment). Note that approximately 50% of all drugs are metabolized at least partly by the p450 enzymes, and 30% of drugs are metabolized primarily by these enzymes. The most important CYP enzymes in drug metabolism are the CYP3A4, CYP2D6 and CYP2C9 enzymes. In accordance with an embodiment of this invention, a separate specific model is employed for one or more of these CYP enzymes.

Figure 2 presents a process flow for a specific way of using a binding affinity model of the present invention in the context of a larger process for predicting whether a particular site on a compound will be metabolized. As shown in Figure 2, a process 201 begins at 203 with receipt of descriptors for the current compound under consideration. In the context of this model, these descriptors are pertinent to the binding of the compound to one or more CYP enzymes (at least CYP2C9 in most embodiments). At 205, a model constructed in accordance with this invention predicts the binding of the current compound to the binding site of interest. Note that the models of this invention may predict a value associated with binding affinity (e.g., Ki) or a simple yes/no (binding or no binding) result. The later case serves as a classification model, and the results have no direct numerical connection to a Ki or pKi value. Regardless of which form the model takes, the process determines at 207

whether the prediction indicates that binding will occur. This acts as a first pass filter for the metabolism model. Assuming that the binding model predicts that the compound under consideration is not, in fact, likely to bind to the binding site of this CYP enzyme, that compound is not considered further. This saves the process from expending additional computational resources on analyzing a compound that will not likely be metabolized by a CYP enzyme under consideration. So, assuming that the process determines that the compound under consideration will not bind (at 207), the process moves on to additional compounds, assuming that such compounds remain to be analyzed. See 209. If additional compounds remain, process control returns to block 203 where the binding descriptors for the next compound under consideration are received. The process then proceeds through operations 205-207 as described above. If no more compounds remain to be considered (i.e., decision 209 is answered in the negative), the process is then completed as illustrated.

Assuming that the compound under consideration has been found to bind sufficiently strongly to the CYP enzyme of interest, process control is directed to 211 where process 201 receives property values relevant to site specific metabolism of the compound under consideration. These property values vary depending upon the form of the metabolism model. If the model employs quantum mechanical analysis, then the relevant properties will include at least an electron distribution about potential reactive sites on the molecule. If, on the other hand, the model is a descriptor-based model, then these property values will be atom or site-specific structural descriptors of the molecule. For quantum mechanical models, the input information requires a detailed three-dimensional structural/electronic representation as described U.S. Patent Application No. 09/258,690 (filed February 26, 1999 and naming De Silva et al., as inventors) and U.S. Patent Application No. 09/613,875 (filed July 10, 2002 and naming Jones et al., as inventors), both incorporated herein by reference. The descriptors required for the second form of model are described in U.S. Patent Application No. 09/811,283 (filed March 15, 2001 and naming Korzekwa et al., as inventors), incorporated herein by reference.

After the relevant input properties have been received, the process must apply the relevant subset of these properties to analyze a particular site on the compound. Thus, as depicted in 201, the next operation involves selecting a particular site on the compound. See 213. With the relevant structural properties for that site at its disposal, the process assesses the reactivity of that site at 215. This assessment is preferably made in accordance with the principals described in one or more of U.S. Patent Application No. 09/258,690, U.S. Patent Application No. 09/613,875, and U.S. Patent Application No. 09/811,283, all previously incorporated by reference. After

the reactivity of the site in question has been ascertained and stored for further consideration, the process determines whether there are additional sites on the compound that require consideration. See 217. If so, process control returns to 213 where the next site is selected and its reactivity is assessed, at 215.

After all sites on the compound under consideration have been analyzed for reactivity (i.e., decision 217 is answered in the negative), the process assesses the metabolic reactivity of the molecule as a whole at 219. It accomplishes this by considering the individual reactivities of the various sites analyzed in operation 215. The reactivity of each site on the molecule contributes to the overall metabolic reactivity of the molecule. Note that the model may employ corrections for accessibility as described in US Patent Application No. 09/902,470 (filed July 9, 2001 and naming Korzekwa et al., as inventors), incorporated herein by reference. After 219, the process determines whether any more compounds remain to be considered at 209. When all compounds have been considered, 209 is answered in the negative and the process is completed as mentioned above.

D. APPLICATIONS AND EXAMPLES

CYP2C9 metabolizes only those compounds that bind to it. As indicated above in the discussion of Figure 2, the CYP binding models of this invention may be used to filter compounds that are unlikely to bind with sufficient strength to one or more CYP enzymes. The computational methodology excludes such compounds from further analysis. Thus, the binding models of this invention are commonly used in conjunction with other software models, such as models that predict actual rates of metabolism for binding compounds and other software that predicts regioselectivity of CYP2C9 for binding compounds. The models of this invention tell a researcher, whether a molecule will bind to CYP2C9. If binds, then it is typically metabolized.

Since compounds can be metabolized by enzymes other than CYP2C9, this model may be used in conjunction with other models that predict binding to non-CYP2C9 isoforms (e.g., CYP3A4, CYP2D6, etc.). Metabolism rates and/or regioselectivity assessments are made by models for those CYP isoforms to which the compound in question is found to bind. Together these various models may be provided as a "suite" of models that predict binding to CYP2C9. The suite may also

include software that predicts overall rates of metabolism and/or regioselectivity in metabolism by CYP enzymes that bind a substrate.

An important application of the present invention is in guided redesign of potential therapeutic compounds that have ADMET/PK difficulties. Using the models of this invention, a skilled chemist or moderately sophisticated software routine can propose modified chemical structures having improved ADMET/PK properties. To do this effectively, the redesign effort may require information on which CYP isoforms metabolize a given compound. The models of this invention provide such information.

Used in conjunction with other models, the CYP binding models of this invention can identify a number of candidate compounds. However, the compounds selected by this invention as likely having a desirable activity may need to be tested *in vitro* or *in vivo*. One of skill in the art will recognize that there are many different ways to experimentally confirm the activity predicted by the invention. Compounds may be tested for predicted ADMET/PK activity by using biochemical assays such as Human Serum Albumin binding, chemical assays such as pK_A and solubility testing, and in vitro biological assays such as metabolism by endoplasmic reticulum fractions of human liver, in order to estimate their actual in vivo ADME/PK properties.

The ability of the present invention to predict the binding affinity of compounds to the CYP2C9 isoform is also applicable in the prediction of drug interactions, as the Ki may be related to the ability of the compound to inhibit metabolism of other drugs that are substrates for CYP2C9. If the compound binds tightly to CYP2C9, then models of this invention can be used to predict/assess drug-drug interactions. In other words, the present invention allows one to predict which substrate bind so tightly that they may be inhibitors to the metabolism of drugs that are normally metabolized by CYP2C9, in large measure.

In a very specific embodiment, the model takes the following form:

pKi = 0.0162471(a_aro) + -0.0237885(a_count) + 0.0470383(b_rotN) + 0.000731773(molecular weight) + 0.00145894(a_heavy) + 0.000744728(PEOE_VSA_HYD) + -0.015514(PEOE_VSA_PNEG) + 0.159486(a_acc) + -0.0111914(a_don) + -0.0253873(a_hyd) + 0.521506(SlogP).

In this expression, the descriptors are defined as follows:

(a_aro) indicates the number of aromatic atoms in a given molecule.

(a_count) is the total number of atoms, and is a measure of the size of the molecule.

(b_rotN) is the number of single, rotatable bonds, and is indicative of the flexibility of the molecule.

(molecular weight) is the total molecular weight of the molecule.

(a_heavy) is the number of non-hydrogen atoms present.

(PEOE_VSA_HYD) refers to the accessible surface of the hydrophobic atoms, or the poorly-charged atoms.

(PEOE_VSA_PNEG) refers to the Van der Waals accessible surface; the partial negative charge is indicative of the number of negatively charged atoms exposed to the surface.

(a_acc) is the number of hydrogen bond acceptor atoms.

(a_don) is the number of hydrogen bond donor atoms, defined as any hydrogen atoms covalently bound to an electronegative atom or group such as an oxygen atom, a sulfur atom, or a nitrogen atom.

(a_hyd) indicates the number of hydrophobic atoms.

(SlogP) is a calculated log of the octanol/water partition coefficient, which includes implicit hydrogens.

This example takes the form of a simple expression of multiple independent variables. More specifically, it is a linear additive model in which the products of coefficients and descriptors are summed. As indicated, many other model forms are possible. They can be expressions, look up tables, docking algorithms, etc. In accordance with this invention, the models must predict binding to a metabolizing enzyme, preferably a CYP enzyme such as CYP2C9. Preferably, they also employ measures of compound lipophilicity, negative charge, and aromatic character as inputs. Examples of the mathematical/logical form of models useful for this invention include linear and non-linear mathematical expressions, neural networks, support vector machines, Bayesian models, classification and regression trees/graphs, clustering approaches, and the like.

E. HARDWARE/SOFTWARE IMPLEMENTATION

Note that many embodiments of this invention are implemented as software and hardware acting under control of particular instructions. Hence, certain embodiments of the present invention employ processes acting or acting under control

of data stored in or transferred through one or more computer systems. Embodiments of the present invention also relate to an apparatus for performing these operations. This apparatus may be specially designed and/or constructed for the required purposes, or it may be a general-purpose computer selectively activated or reconfigured by a computer program and/or data structure stored in the computer. The processes presented herein are not inherently related to any particular computer or other apparatus. In particular, various general-purpose machines may be used with programs written in accordance with the teachings herein, or it may be more convenient to construct a more specialized apparatus to perform the required method steps.

In addition, embodiments of the present invention relate to computer readable media or computer program products that include program instructions and/or data (including data structures) for performing various computer-implemented operations. Examples of computer-readable media include, but are not limited to, magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD-ROM devices and holographic devices; magneto-optical media; semiconductor memory devices, and hardware devices that are specially configured to store and perform program instructions, such as read-only memory devices (ROM) and random access memory (RAM), and sometimes application-specific integrated circuits (ASICs), programmable logic devices (PLDs) and signal transmission media for delivering computer-readable instructions, such as local area networks, wide area networks, and the Internet. The data and program instructions of this invention may also be embodied on a carrier wave or other transport medium (e.g., optical lines, electrical lines, and/or airwaves). Examples of program instructions include both machine code, such as produced by a compiler, and files containing higher level code that may be executed by the computer using an interpreter.

Figure 3 is a schematic illustration of an Internet-based embodiment of the current invention. See 300. According to a specific embodiment, a client 302, at a drug discovery site, for example, sends data 308 identifying organic molecules 308 to a processing server, 306 via the Internet 304. The organic molecules are simply the molecules that the client wishes to have analyzed by the current invention. At the processing server 306, the molecules of interest are analyzed by a model 312, which predicts particular ADMET/PK properties (including binding affinity for CYP2C9), for example. The processing server may also redesign compounds to modify their CYP binding or activity.

After the analysis, the predicted CYP binding 310 (and any other appropriate information) are sent via the Internet 304 back to the client 302. Many general purpose and specialized computer systems are suitable as either the client 302 or the processing server 306. In a specific embodiment, standard transmission protocols such as TCP/IP (transmission control protocol/internet protocol) are used to communicate between the client 302 and processing server 306. Security measures such as SSL (secure socket layer), VPN (virtual private network) and encryption methods (e.g., public key encryption) can also be used.

F. OTHER EMBODIMENTS

Although the above has generally described the present invention according to specific processes and apparatus, the present invention has a much broader range of applicability. In particular, the present invention is not limited to a particular class of descriptor or compound. Of course, one of ordinary skill in the art would recognize other variations, modifications, and alternatives.

CLAIMS

what is claimed is

1. A computer-implemented method of predicting binding of a compound to the 2C9 isoform of the cytochrome p450 family of enzymes (CYP2C9), the method comprising:

(a) receiving a value representing the lipophilicity of the compound;

(b) receiving a value representing the partial negative charge or negative charge associated with the compound;

(c) receiving a value representing the aromatic character of the compound; and

(d) calculating the binding of the compound to CYP2C9 by providing the values received in (a)-(c) to an expression treating said values as independent variables and treating the binding to CYP2C9 as a dependent variable.

2. The method of claim 1, wherein the expression provides a linear relationship between said values and the binding to CYP2C9.

3. The method of claim 1, wherein the expression was derived by performing a regression on a data set comprising values of lipophilicity, partial negative charge, aromatic character, and binding to CYP2C9 for various compounds comprising a training set.

4. The method of claim 1, wherein calculating the binding of the compound to CYP2C9 comprises calculating a value of Ki or pKi for the compound-CYP2C9 binding.

5. The method of claim 1, wherein the value representing the lipophilicity comprises a measured or predicted value of a partitioning property of the compound.

6. The method of claim 5, wherein the partitioning property comprises a partition coefficient or distribution coefficient of the compound.

7. The method of claim 1, wherein the value representing the lipophilicity comprises a measure or prediction of the hydrophobicity of the compound.

8. The method of claim 1, wherein the value representing the lipophilicity comprises the number of hydrophobic atoms in the compound or the surface area of the compound occupied by hydrophobic atoms.

9. The method of claim 1, wherein the value representing the partial negative charge on the compound comprises the surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

10. The method of claim 1, wherein the value representing the aromatic character of the compound comprises one or more of (i) the number of aromatic atoms in the compound, (ii) the type of aromatic atoms in the compound, (iii) the number of aromatic rings in the compound, and (iv) the type of aromatic rings in the compound.

11. The method of claim 1, further comprising:

receiving a value representing the compound size,

wherein the expression further treats the value representing size as an independent variable.

12. The method of claim 11, wherein the value representing compound size comprises at least one of molecular weight, number of atoms, number of atoms greater than a particular atomic weight, and molecular surface area.

13. The method of claim 1, further comprising:

receiving a value representing the compound flexibility,

wherein the expression further treats the value representing flexibility as an independent variable.

14. The method of claim 13, wherein the value representing compound flexibility comprises at least one of number of rotatable bonds, and fraction of bonds that are rotatable.

15. The method of claim 1, further comprising:

receiving a value representing compound flexibility; and

receiving a value representing compound size,

wherein the expression further treats the values representing compound flexibility and compound size as independent variables.

16. The method of claim 15, wherein the expression includes at least the following independent variables: number of aromatic atoms, molecular weight, number of rotatable bonds, a partitioning property, and a surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

17. The method of claim 1, further comprising:

performing (a)-(d) for a plurality of different compounds; and

selecting those compounds for which the calculated binding to CYP2C9 exceeds a predefined value; and

characterizing those selected compounds as potential inhibitors having drug interference difficulties.

18. The method of claim 1, further comprising:

performing (a)-(d) for a plurality of different compounds; and

selecting those compounds for which the calculated binding to CYP2C9 exceeds a predefined value; and

predicting reactivity of sites on those selected compounds using a computer model of reactivity.

19. A computer program product comprising a machine readable medium on which is provided instructions for predicting binding of a compound to the 2C9 isoform of the cytochrome p450 family of enzymes (CYP2C9), wherein the instructions comprise:

(a) code for receiving a value representing the lipophilicity of the compound;

(b) code for receiving a value representing the partial negative charge or negative charge associated with the compound;

(c) code for receiving a value representing the aromatic character of the compound; and

(d) code calculating the binding of the compound to CYP2C9 by providing the values received in (a)-(c) to an expression treating said values as independent variables and treating the binding to CYP2C9 as a dependent variable.

20. The computer program product of claim 19, wherein the expression provides a linear relationship between said values and the binding to CYP2C9.

21. The computer program product of claim 19, wherein the code for calculating the binding of the compound to CYP2C9 comprises code for calculating a value of Ki or pKi for the compound-CYP2C9 binding.

22. The computer program product of claim 19, wherein the value representing the lipophilicity comprises a partition coefficient or distribution coefficient of the compound.

23. The computer program product of claim 19, wherein the value representing the lipophilicity comprises the number of hydrophobic atoms in the compound or the surface area of the compound occupied by hydrophobic atoms.

24. The computer program product of claim 19, wherein the value representing the partial negative charge on the compound comprises the surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

25. The computer program product of claim 19, wherein the value representing the aromatic character of the compound comprises one or more of (i) the number of aromatic atoms in the compound, (ii) the type of aromatic atoms in the compound, (iii) the number of aromatic rings in the compound, and (iv) the type of aromatic rings in the compound.

26. The computer program product of claim 19, wherein the instructions further comprise:

code for receiving a value representing compound flexibility; and

code receiving a value representing compound size,

wherein the expression further treats the values representing compound flexibility and compound size as independent variables.

27. The computer program product of claim 26, wherein the expression includes at least the following independent variables: number of aromatic atoms, molecular weight, number of rotatable bonds, a partition coefficient, and a surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

28. The computer program product of claim 19, wherein the instructions further comprise:

code for performing (a)-(d) for a plurality of different compounds; and

code for selecting those compounds for which the calculated binding to CYP2C9 exceeds a predefined value; and

code for (i) characterizing those selected compounds as potential inhibitors having drug interference difficulties or (ii) predicting reactivity of sites on those selected compounds using a computer model of reactivity.

29. A computer-implemented method of creating a multivariate model for predicting the binding of compounds to the 2C9 isoform of the cytochrome p450 family of enzymes (CYP2C9), the method comprising:

(a) for each compound for a plurality of compounds in a training set, receiving values representing the binding of the compound to CYP2C9, the lipophilicity of the compound, the partial negative charge or negative charge associated with the compound, and the aromatic character of the compound; and

(b) fitting the values to create the multivariate model of binding to CYP2C9 as a function of lipophilicity, partial negative charge or negative charge, and aromatic character.

30. The method of claim 29, wherein the model comprises an expression providing a linear relationship between said values and the binding to CYP2C9.

31. The method of claim 29, wherein (b) comprises performing a regression on a on the values for the compounds comprising a training set.

32. The method of claim 29, wherein the values representing the binding of the compound to CYP2C9 comprise values of Ki or pKi.

33. The method of claim 29, wherein the values representing the lipophilicity comprise measured or predicted values of a partitioning property of the compound.

34. The method of claim 33, wherein the partitioning property comprises a partition coefficient or distribution coefficient of the compound.

35. The method of claim 29, wherein the values representing the lipophilicity comprise the number of hydrophobic atoms in the compound or the surface area of the compound occupied by hydrophobic atoms.

36. The method of claim 29, wherein the values representing the partial negative charge on the compound comprise the surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

37. The method of claim 29, wherein the values representing the aromatic character of the compound comprise one or more of (i) the number of aromatic atoms in the compound, (ii) the type of aromatic atoms in the compound, (iii) the number of aromatic rings in the compound, and (iv) the type of aromatic rings in the compound.

38. The method of claim 29, wherein (a) further comprises receiving a value representing the compound size, and wherein the multivariate model created in (b) is a function of compound size also.

39. The method of claim 29, wherein (a) further comprises receiving a value representing the compound flexibility, and wherein the multivariate model created in (b) is a function of compound flexibility also.

40. The method of claim 29, wherein (a) further comprises

receiving a value representing compound flexibility; and

receiving a value representing compound size,

and wherein the multivariate model created in (b) is a function of compound flexibility and compound size also.

41. The method of claim 40, wherein the multivariate model created in (b) is an expression relating the binding to CYP2C9 to at least the following independent variables: number of aromatic atoms, molecular weight, number of rotatable bonds, a partitioning property, and a surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

42. A computer program product comprising a machine readable medium on which is stored program instructions for creating a multivariate model for predicting the binding of compounds to the 2C9 isoform of the cytochrome p450 family of enzymes (CYP2C9), wherein the program instructions comprise:

(a) code for receiving values, for each compound for a plurality of compounds in a training set, representing the binding of the compound to CYP2C9, the lipophilicity of the compound, the partial negative charge or negative charge associated with the compound, and the aromatic character of the compound; and

(b) code for fitting the values to create the multivariate model of binding to CYP2C9 as a function of lipophilicity, partial negative charge or negative charge, and aromatic character.

43. The computer program product of claim 42, wherein the model comprises an expression providing a linear relationship between said values and the binding to CYP2C9.

44. The computer program product of claim 42, wherein (b) comprises code for performing a regression on a on the values for the compounds comprising a training set.

45. The computer program product of claim 42, wherein the values representing the binding of the compound to CYP2C9 comprise values of Ki or pKi.

46. The computer program product of claim 42, wherein the lipophilicity comprises one or more of a partition coefficient of the compound, a distribution coefficient of the compound, the number of hydrophobic atoms in the compound, or the surface area of the compound occupied by hydrophobic atoms.

47. The computer program product of claim 42, wherein the values representing the partial negative charge on the compound comprise the surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

48. The computer program product of claim 42, wherein the values representing the aromatic character of the compound comprise one or more of (i) the number of aromatic atoms in the compound, (ii) the type of aromatic atoms in the compound, (iii) the number of aromatic rings in the compound, and (iv) the type of aromatic rings in the compound.

49. The computer program product of claim 42, wherein (a) further comprises

code for receiving a value representing compound flexibility; and

code for receiving a value representing compound size,

and wherein the multivariate model created in (b) is a function of compound flexibility and compound size also.

50. The computer program product of claim 49, wherein the multivariate model created with (b) is an expression relating the binding to CYP2C9 to at least the following independent variables: number of aromatic atoms, molecular weight, number of rotatable bonds, a partition coefficient, and a surface area of the compound occupied by atoms having a partial charge of magnitude greater than a predefined level.

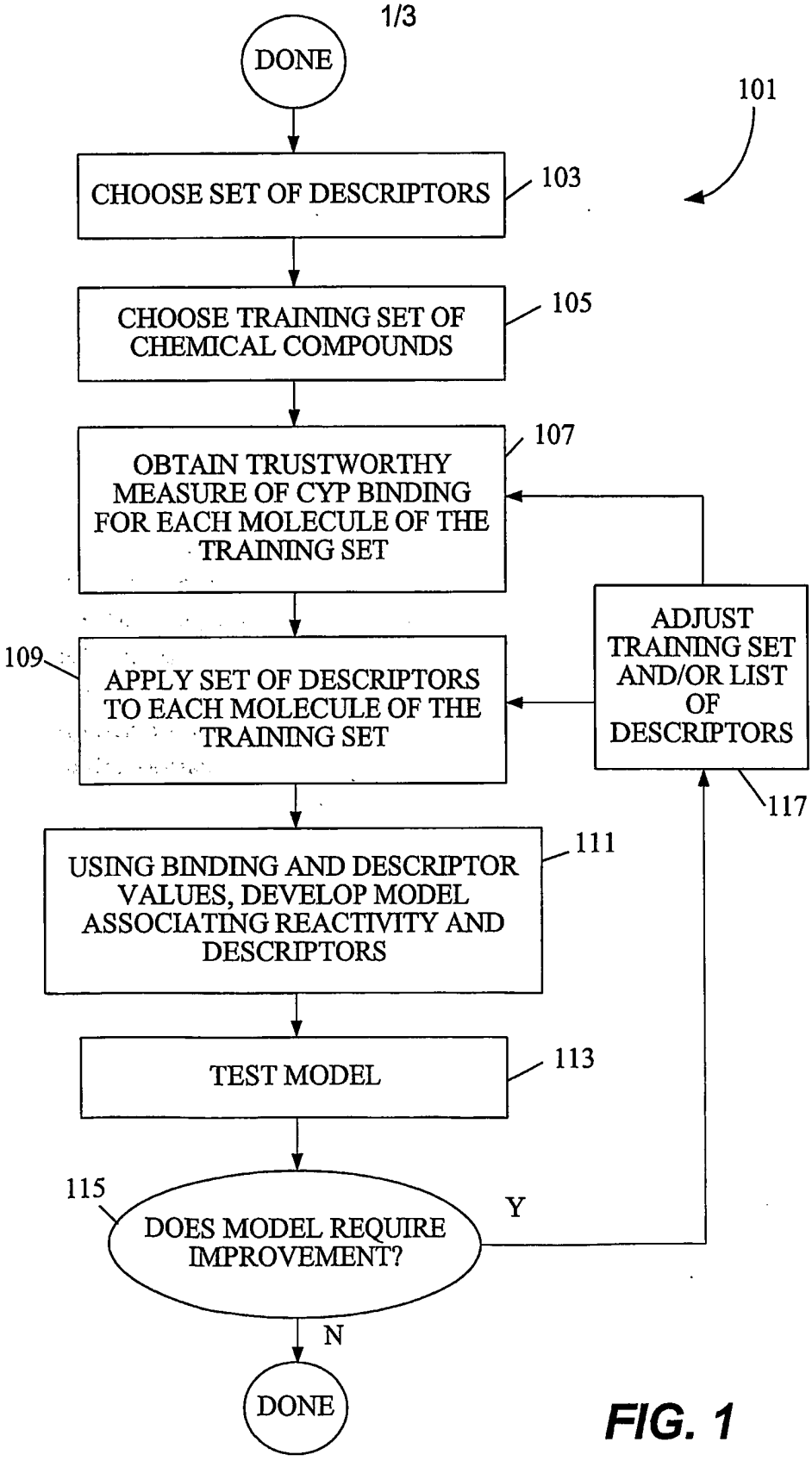

DONE
101
CHOOSE SET OF DESCRIPTORS
103
CHOOSE TRAINING SET OF CHEMICAL COMPOUNDS
105
OBTAIN TRUSTWORTHY MEASURE OF CYP BINDING FOR EACH MOLECULE OF THE TRAINING SET
107
109
APPLY SET OF DESCRIPTORS TO EACH MOLECULE OF THE TRAINING SET
ADJUST TRAINING SET AND/OR LIST OF DESCRIPTORS
117
USING BINDING AND DESCRIPTOR VALUES, DEVELOP MODEL ASSOCIATING REACTIVITY AND DESCRIPTORS
111
TEST MODEL
113
115
DOES MODEL REQUIRE IMPROVEMENT?
Y
N
DONE

FIG. 1

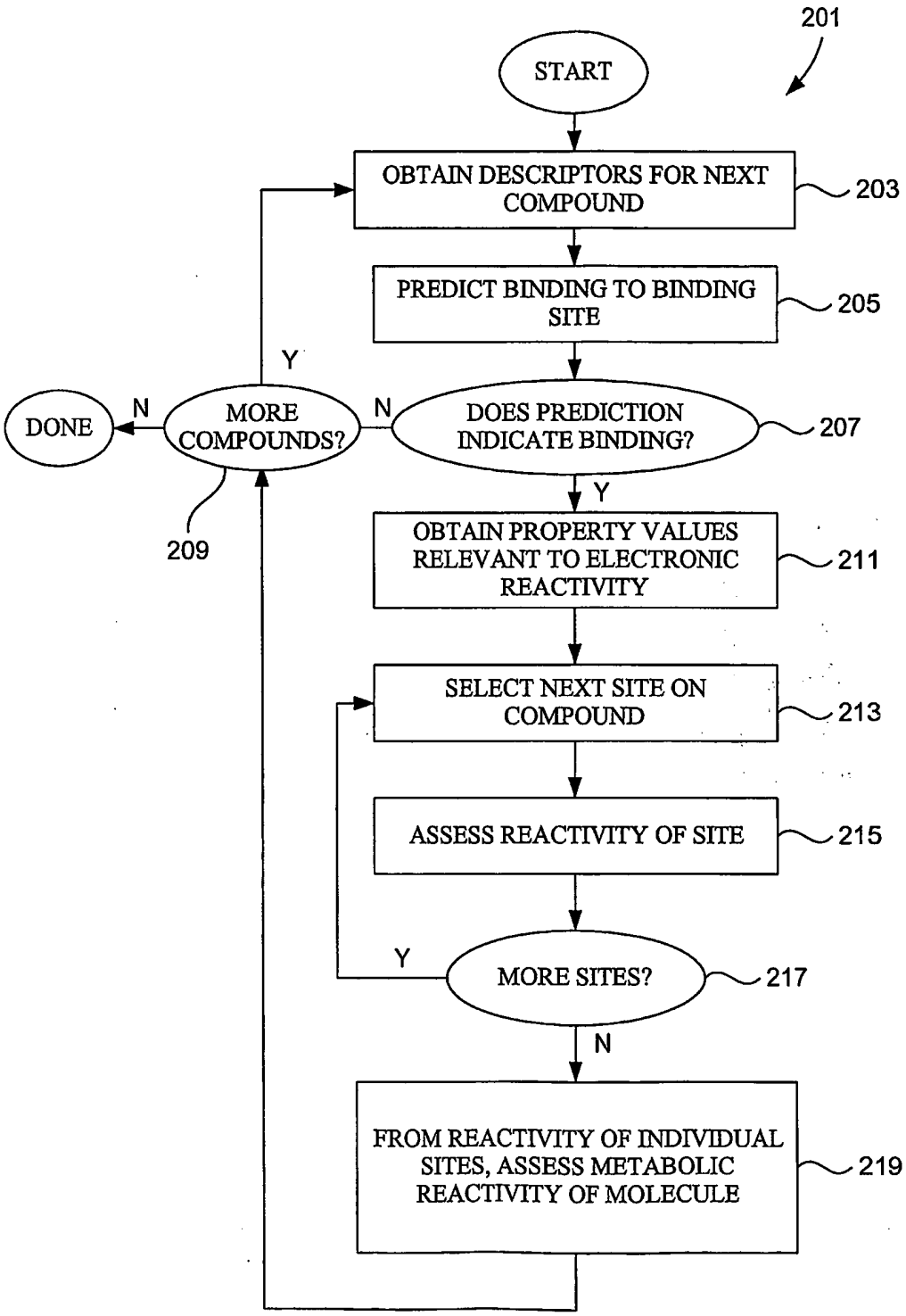

201
START
OBTAIN DESCRIPTORS FOR NEXT COMPOUND
203
PREDICT BINDING TO BINDING SITE
205
DOES PREDICTION INDICATE BINDING?
207
MORE COMPOUNDS?
DONE
Y
N
N
Y
209
OBTAIN PROPERTY VALUES RELEVANT TO ELECTRONIC REACTIVITY
211
SELECT NEXT SITE ON COMPOUND
213
ASSESS REACTIVITY OF SITE
215
MORE SITES?
217
Y
N
FROM REACTIVITY OF INDIVIDUAL SITES, ASSESS METABOLIC REACTIVITY OF MOLECULE
219

FIGURE 2

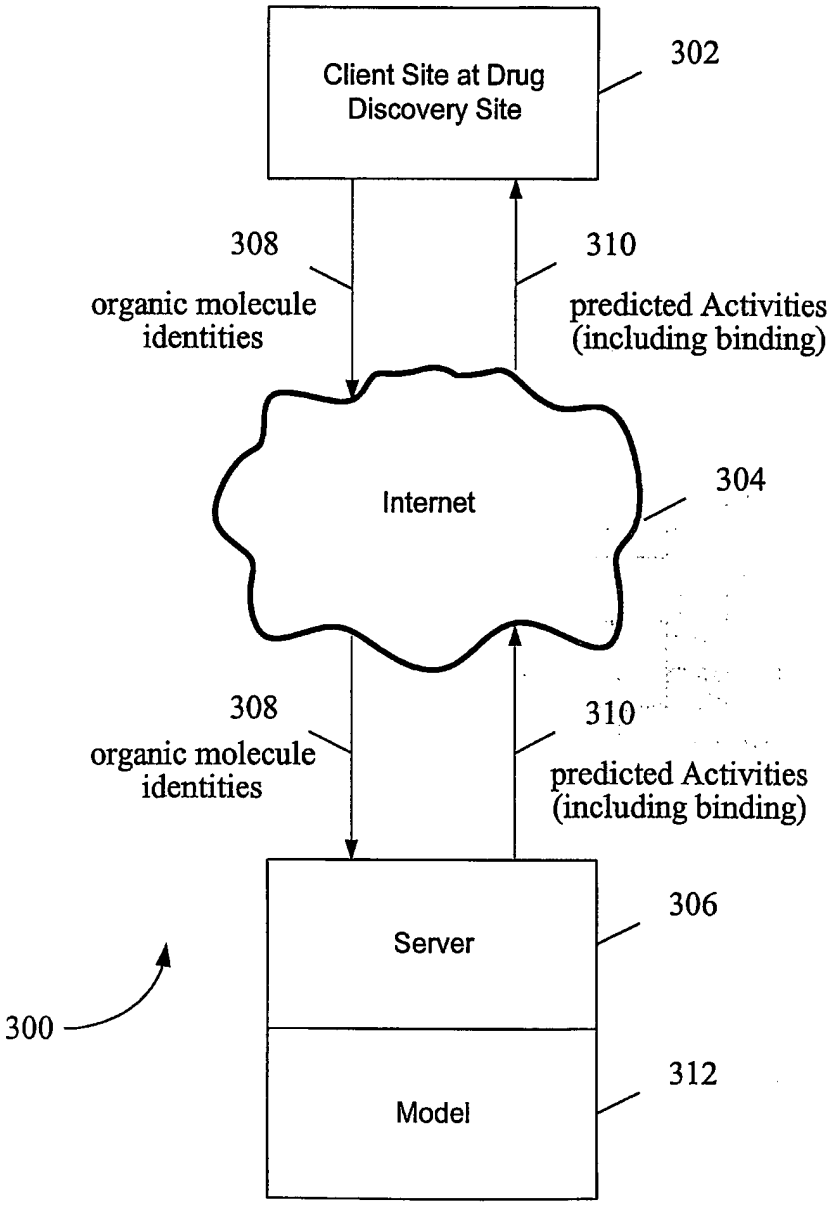

Client Site at Drug Discovery Site
302
308
organic molecule identities
310
predicted Activities (including binding)
Internet
304
308
organic molecule identities
310
predicted Activities (including binding)
Server
306
Model
312
300

FIGURE 3